

Praktikum z ekonometrie

VŠE Praha

Tomáš Formánek

Block 1 – Missing data – Outline

- 1 The nature of missing data
- 2 Traditional treatment of missing data
- 3 Modern Approaches to missing data
 - Multiple imputation for CS data
 - Imputation for TS data
- 4 Missing dependent variable data

The nature of missing data

Missing completely at random (*MCAR*)

- The probability that an observation X_i is missing is unrelated to the value of X_i or to the value of any other variables.
- Any piece of data is equally likely to be missing.
- Analyses based on data with *MCAR* observations remain unbiased. We may lose power (increased standard errors), but the estimated parameters are not biased by the absence of data.

Missing at random (*MAR*)

- Data meets the requirement that missingness does not depend on the value of X_i after controlling for other variable(s).
- For example, data are *MCAR* for any demographic category (but the probability of missingness differs among such groups).

Missing Not at Random (*MNAR*)

- Missingness of X_i depends on its value (e.g. income in surveys)
- The only way to obtain unbiased estimates of (regression) parameters is to model the missingness.

Traditional treatment of missing data

Listwise deletion (complete cases analysis)

- We omit all rows with missing data – missing information for at least one variable in the i -th individual observation. Then, we run our analyses on the observations that remain. This often results in a substantial decrease in sample size. Under the assumption that data are missing completely at random, LRMs lead to unbiased parameter estimates – still, we lose power due to exclusion of (potentially large number of) observations.

R code

```
newData <- data[complete.cases(data)==T, ]  
# data is a data.frame  
# or  
newData <- na.omit(data)
```

Hot deck imputation

- Historically used by the US Census Bureau (since 1950's). Respondent's missing data were replaced by observed replacement data – drawn at random from a group of similar participants. Suitable, given only a few missing observations need to be replaced and given the draw is random.

Traditional treatment of missing data

Mean substitution

- ✓ Simple
- ✗ In simple linear regression models (SLRMs), this adds no new information but increases sample size – that leads to underestimated standard errors only.

Example: Data on salary and citation level of publications. 62 cases with complete data and 7 cases for which the citation index was missing. Correlations and regression coefficients were compared as follows:

Analysis	n	$corr$	$\hat{\beta}_1$	$s.e.(\hat{\beta}_1)$
Complete cases only	62	.55	310.747	60.95
With mean substitution	69	.54	310.747	59.12

Mean substitution, contnd.

- Mean imputation can be usefull for multiple linear regression models, especially when data are missing as MCAR.
- It is fast, simple, easy to implement, and no cases are excluded.
- Even under MCAR, this method still leads to underestimation of coefficient variance.
- Bias in variance estimation is proportional to $(\text{nobs} - 1)/(\text{nobs} + \text{nmis} - 1)$.

Smaller standard errors increase the possibility of Type I error (rejecting true null hypothesis).

Regression substitution

- Uses linear regression (auxiliary LRM) to predict what the missing values of regressors should be – on the basis of other variables that are present.
- May be useful for MLRMs.
- For SLRMs, this approach would be equivalent to mean substitution. We do not add more information but we increase the sample size and (spuriously) reduce standard errors.

Stochastic regression substitution

- Build on regression substitution: this approach adds a randomly sampled residual term from the normal (or other) distribution to each value estimated by regression substitution. Adding a bit of random error to each substitution reduces, but does not eliminate, the problem of spurious reduction of the standard errors.

Maximum Likelihood Expectation-Maximization

- Maximum likelihood approach – alternative to OLS – for the estimation of missing values.

Many approaches exist (e.g. the Expectation-Maximization algorithm)

https://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing-Part-Two.html

Predictive mean matching (PMM)

- Discussed next, within the Multiple imputation section

Multiple Imputation (MI)

R: {mice}, {mi}, {Amelia}, ...

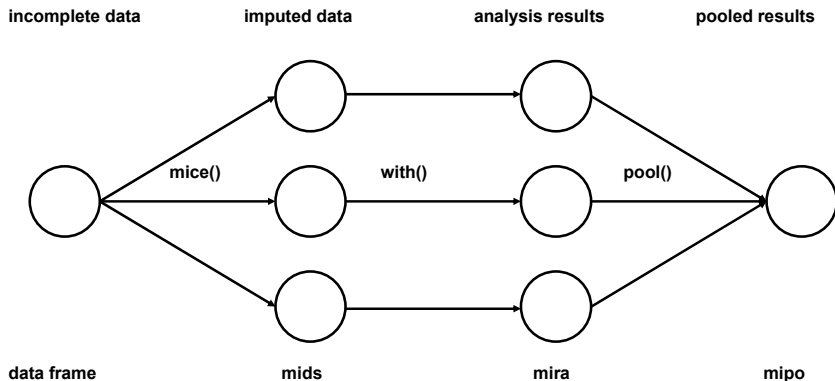
MI motivation and algorithm

- Use PMM to create several (say, 5) – imputed values for each missing item – regressor x_{ij} .
- Each version of imputed data is organized into a separate data set and used for estimation (OLS, ML or other adequate approach).
- Information obtained from all estimates is conveniently summarized.

Multiple imputation (MI)

Multiple imputation scheme (example with $m = 3$ imputations):

- The first step – `mice()` function – involves PMM.



`{mice}` object-types used for MI.

Multiple imputation (MI)

Multiple imputation - 7 choices to be made

- 1 Decide on MAR assumption plausibility (MAR/MNAR).
- 2 Imputation model choice (univariate, multivariate, data type).
- 3 Choice of predictors for MI.
- 4 Should we impute variables that are functions of incomplete variables (e.g. interaction terms)?
- 5 If data is missing in more than one variable, ordering for imputation can affect results.
- 6 MI is based on a numerical algorithm (say, PMM): we need to choose starting setup (“proximity” conditions and possibly other hyper-parameters, say number of iterations).
- 7 We need to choose m – the number of imputed datasets.

In R (`{mice}`), most of the choices have generally valid default setting. However, the choices are made & affect the resulting imputations.

Predictive mean matching (pmm) in R – general description

- Implemented in {mice} and other packages.
- General purpose semi-parametric imputation method.
- Suitable especially for imputing quantitative variables that are not normally distributed.
- Imputations are restricted to previously observed values.
- Can preserve non-linear relations even if the structural part of the imputation model is wrong.

Multiple imputation (MI)

Predictive mean matching (pmm) in R – algorithm

Suppose there is a single variable x that has some cases with missing data, and a set of relevant variables \mathbf{z} with no missing data:

$$\{x_i, z_{i1}, \dots, z_{iL}\}; \quad i = 1, \dots, n$$

- 1 For cases with no missing x_i data, estimate LRM $x_i \leftarrow \mathbf{z}_i$, producing $\hat{\beta}$ and $\text{var}(\hat{\beta})$ estimates.
- 2 Make a random draw from the “posterior predictive distribution” of $\hat{\beta}$, producing a new set of coefficients $\hat{\beta}^*$.
 - Random draw from a multivariate normal distribution with mean $\hat{\beta}$ and cov. matrix $\text{var}(\hat{\beta})$.
Other distributions can be used, upon data and model used.
 - This step is necessary to produce sufficient variability in the imputed values, and is common to all “proper” methods of MI.

Multiple imputation (MI)

Predictive mean matching (pmm) in R – algorithm contnd.

- 3 Using $(z_i \hat{\beta}^*)$, generate predicted values \hat{x}_i for **all cases**, both with data missing on x_i and with data observations present.
- 4 For each case with missing x_i , use the fitted \hat{x}_i and search for a set of similar “closely matching” \hat{x}_ℓ predictions (here, ℓ is a row index, different from i ; we are only interested in ℓ cases where x_ℓ is observed).
 - Similarity (“closeness”) rules are defined separately.
- 5 For each missing x_i , randomly choose one x_ℓ (observation) from the set of “close” observations and use its observed value to substitute for the missing value (x_ℓ observation is imputed, NOT the \hat{x}_ℓ value).
- 6 For MI, repeat steps 2 through 5 m -times to produce m imputed datasets.

Multiple imputation (MI)

Predictive mean matching (pmm) in R – recap.

- Compared with regression-based methods, PMM produces imputed values that are much more like real values.
 - If the original variable is skewed, imputed values will also be skewed.
 - If the original variable is bounded by 0 and 100, imputed values will also be bounded by 0 and 100.
 - If the real values are discrete (say, number of children), imputed values will also be discrete.
- Generally speaking, there's no mathematical proof/theory to “justify” PMM (efficiency).
- PMM efficiency can be demonstrated by Monte Carlo simulations.

Multiple imputation (MI)

Multiple Imputation (empirical output example)

Regression coefficients from five imputed data sets

Data set	Estimated parameter	b_0	b_1	b_2	b_3	b_4	b_5
1	Coefficient	-11.535	-2.780	1.029	-.031	-0.359	0.572
	Variance	43.204	3.323	0.013	0.013	0.013	0.012
2	Coefficient	-11.501	-4.149	1.040	-0.093	-0.583	0.876
	Variance	40.488	2.680	0.010	0.009	0.009	0.007
3	Coefficient	-10.141	-5.038	0.766	0.123	-0.252	0.625
	Variance	42.055	3.301	0.010	0.010	0.010	0.009
4	Coefficient	-11.533	-6.920	0.870	0.084	-0.458	0.815
	Variance	28.751	1.796	0.081	0.007	0.007	0.007
5	Coefficient	-14.586	-1.115	0.718	0.050	-0.373	0.814
	Variance	32.856	2.362	0.009	0.009	0.009	0.008
	Mean b_i	-11.859	-4.000	0.885	0.027	-0.405	0.740
	Mean Var. (\bar{W})	37.471	2.692	0.025	0.010	0.010	0.009
	Var. of b_i (B)	2.682	4.859	0.022	0.008	0.015	0.018
	T						
	\sqrt{T}	40.69	8.523	0.051	0.020	0.028	0.031
	t	6.379	2.919	0.226	0.141	0.167	0.176
		-1.859	-1.370	3.916*	0.191	2.425*	4.204*

* $p < .05$ "Var." refers to the squared standard error of the coefficient.

https://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing-Part-Two.html

- Univariate TS imputation
 - R packages `imputeTS`, `zoo`, etc.
 - LOCF, linear & spline interpolation, Kalman filter, ...
- Multivariate TS imputation
 - R package `Amelia`
 - Use time trend (and polynomes), leads, lags, priors, ...

Special considerations apply to missing dependent variable data

- If we can assume that data are missing completely at random (*MCAR*), we will lose power because of smaller sample sizes, but we will not have problems with biased estimates.
- If data are missing not at random (*MNAR*), the **only way to obtain an unbiased estimate of parameters is to model missingness**. In other words, we need to use a model that accounts for the missing data.
- Broadly speaking, such models are:
 - Censored Regression Models (e.g. duration analysis)
 - Truncated Regression Models
 - Sample Selection Correction models (Heckit)
 - ...